RECONNAISSANCE DE MOTS MANUSCRITS PAR MODELES MARKOVIENS

T. PAQUET, C. OLIVIER, M. AVILA, Y. LECOURTIER

La3i, Université de Rouen, U.F.R. des sciences
76821 Mont Saint Aignan cédex, France.

RÉSUMÉ

Nous présentons dans cette communication une méthode de reconnaissance de mots manuscrits cursifs d'un petit lexique. Nous évitons d'utiliser un modèle explicite en lettres séparées de façon à contourner le problème délicat de la segmentation et la complexité des modèles cherchant à recombiner les meilleurs possibilités de segmentation. Nous utilisons deux représentations structurelles l'une en traits l'autre en graphèmes, qui sont toutes deux modèlisées par des processus de Markov. Nous discutons du choix de l'ordre des modèles utilisés à l'aide d'un critère d'information d'Akaïke. La probabilité du modèle considéré conditionellement à la séquence observée est évaluée en prenant en compte la longueur de cette séquence. Nous présentons des résultats de reconnaissance sur une base d'image de chèques.

ABSTRACT

This paper deals with the global recognition of a small lexicon of words based on a pseudo-segmentation stage introducng anchor points. We avoid the difficult problem of finding the segmentation of the word into letters and the complexity involved by such models to build the possible letter graphes. We use two structural representation of the word, strokes or graphemes, each of them being analysed using a Markov Model. These simple models are individualy optimised by a rigorous choice of there order to fit the structural properties of the observed data using Akaïke information criteria. The conditionnal probability to have word model given the observation sequence is computed by taking into account the length of the sequence. Results of the study are presented on French check images.

1. INTRODUCTION

Les nombreuses études menées au cours des trente dernières années dans le domaine de la reconnaissance automatique de l'écriture manuscrite [1],[2], montrent combien il est difficile de modèliser l'écriture cursive de façon à s'affranchir d'un style d'écriture particulier. Les modèles de Markov cachés ont été employés avec succès pour la reconnaissance de la parole [3] et plus récemment pour la reconnaissance de l'écriture manuscrite. Ceci est principalement dû à la puissance de modèlisation de la variabilité d'enchaînement des lettres offerte par de tels modèles [4][5][6]. Nous présentons plus particulièrement dans cet article une technique d'optimisation de modèles stochastiques appliqués à la reconnaissance de l'écriture manuscrite cursive. Trois stratégies de reconnaissances peuvent être développées pour la reconnaissance de mots. La première procède à la reconnaissance des lettres qui composent le mot manuscrit et nécessite une étape préalable de segmentation du mot. La seconde consiste à reconnaitre le mot dans sa globalité sans tenter de le segmenter en lettres. La troisième approche, qui est celle envisagée dans cet article, est une approche globale se référant à une description analytique du mot basée sur des primitives structurelles et une pseudo-segmentation du mot [7]. De cette façon, on contourne le délicat problème de la segmentation de l'écriture cursive en lettres, et la complexité des traitements à appliquer pour proposer des suites de lettres cohérentes. Deux représentations structurelles sont envisagées pour effectuer la reconnaissance s'appuyant, soit sur une description du mot à l'aide de traits élémentaires, soit sur une description en graphèmes (fragments d'écriture cursive). Chacune de ces descriptions est modèlisée par un modèle de Markov dont les états sont soit des traits soit des graphèmes. Ces modèles sont optimisés par un choix rigoureux de leur ordre en utilisant un critère d'information du type critère d'Akaïke [8]. Ce type de critère est très souvent utilisé en traitement du signal pour déterminer l'ordre de modèles AR ou ARMA, et également en théorie de l'information pour optimiser la longueur de chaînes de données selon le principe de description de longueur minimale [9].

Nous présentons tout d'abord la nature de la représentation structurelle des mots utilisée. Nous présentons ensuite les modèles stochastiques utilisés et leur optimisation. Enfin nous exposons les résultats de reconnaissance obtenus par l'application de la méthode présentée sur une base de données significative.

2. MODELE STRUCTUREL

L' étude en cours concerne la reconnaissance d'un vocabulaire limité de mots, le vocabulaire de libellés d'un montant manuscrit présent sur un chèque bancaire, soit 27 mots . Nous disposons de deux bases d'images binarisées de montants manuscrits provenant du Service de Recherche Technique de la Poste (SRTP) de 950 montants soit 3600 mots pour chacune des bases environ. Les pré-traitements destinés à éliminer les symboles pré-imprimés sur la zone d'écriture tels que les lignes d'appui et les deux barres parallèles ont été effectués par les services de la poste. Chaque ligne de texte est localisée en déterminant les lignes d'appui supérieure et inférieure de

l'écriture déterminant ainsi la zone du corps des lettres minuscules. Nous renvoyons à [7] pour une description des traitements effectués qui permettent de localiser chacun des mots du montant manuscrit (fig. 1). L'analyse de la phrase s'effectue par modèlisation Markovienne où les états du modèle sont les mots. Le travail présenté dans la présente communication concerne la reconnaissance de mots isolés et fait suite aux travaux présentés dans [10].

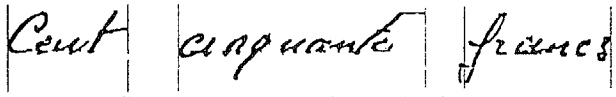

figure 1 : segmentation de la phrase.

Un modèle structurel de chaque mot segmenté est construit en se référant à l'axe médian du mot. Celui-ci est déterminé par la droite de régression déterminée sur les points dans la zone du corps des minuscules. Les points d'intersection du squelette du mot avec l'axe médian sont les points d'ancrage ou germes sur lesquels s'appuie la description structurelle du mot utilisée (fig. 2). Deux types de représentations du mot sont construites autour de l'axe médian.

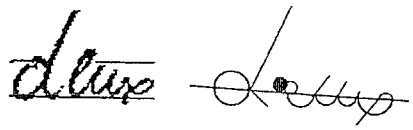

figure 2 : type de représentation structurelle utilisée.

2.1. Description en traits élémentaires

Les éléments de cette description sont les traits de longueur minimale ayant leurs extrémités sur un point d'ancrage. Un ensemble de 8 traits élémentaires a été défini pour cette étude (fig. 3). Cet ensemble réduit ne permet qu'une description grossière du mot puisqu'un trait se terminant dans la zone médiane d'écriture supérieure sera étiqueté trait supérieur quelle que soit son inclinaison ou sa courbure. En limitant volontairement la complexité de la description nous espérons obtenir une description plus robuste aux déformations du tracé induites par la grande variabilité des styles d'écritures.

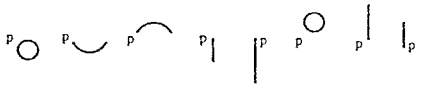

figure 3 : alphabet de huit traits élémentaires.

2.2. Description en graphèmes

La seconde description utilisée traduit davantage les entitées cursives du mot. S'il est souhaitable de se rapprocher le plus possible d'une description du mot cursif en lettres, nous savons qu'un tel modèle suppose une segmentation parfaite du mot en lettres. Cette propriété n'étant jamais atteinte en pratique, nous avons choisi d'utiliser une description en fragments cursifs obtenue en regroupant les traits élémentaires autour des points d'ancrage. Chaque point d'ancrage associé aux traits s'y rattachant constitue un graphème ou fragment d'écriture cursive. De ce fait, une telle description n'est envisageable que sur un vocabulaire limité puisque nous perdons l'universalité de la description en lettres. Nous espérons cependant rester plus proche de la réalité cursive du mot et en conséquence garantir une robustesse de la description aux variabilités de styles d'écritures.

Cette représentation engendre cependant une difficulté: à l'instar du modèle lettres qui se décompose sur un alphabet de 26 symboles, nous devons être capables de définir un alphabet de graphèmes permettant de garantir une description suffisamment riche pour permettre la reconnaissance des mots en présence. Nous devons donc tenter de déterminer un alphabet de graphèmes de façon non supervisée.

2.3. Sélection d'un alphabet de graphèmes

Nous proposons ici une première approche permettant la construction d'un alphabet de graphèmes de façon non supervisée, c'est à dire sans aucune connaissance sur les classes de graphèmes en présence. Parmi les nombreuses méthodes de reconnaissance automatique de données exposées dans la littérature [11], nous avons retenu un algorithme à seuils permettant de sélectionner un alphabet de cardinal inconnu. La première étape de la phase de sélection consiste à répertorier l'ensemble des situations rencontrées sur la base d'apprentissage avec leur fréquence d'apparition respective. Les graphèmes les plus fréquents sont alors retenus pour constituer un alphabet de base. La seconde étape est destinée à l'analyse des graphèmes non encore sélectionnés. Nous utilisons une distance de Hamming afin d'évaluer la ressemblance entre les graphèmes non sélectionnés et les éléments de l'alphabet de base. Un graphème est agrégé à un élément de l'alphabet de base si leur distance est inférieure à un seuil déterminé T. Dans le cas contraire, un graphème non encore sélectionné dont la distance à chacun des graphèmes de l'alphabet de base est supérieure à T constitue alors un nouvel élément de l'alphabet. Nous obtenons un alphabet dont chaque élément est constitué d'une configuration minimale de traits devant être impérativement présents ou absents dans le graphème et de traits optionnels provenant de l'étape de fusion (fig. 4). Le cardinal de l'alphabet ainsi sélectionné dépend directement des valeurs des seuils utilisés. Un ensemble de 42 graphèmes permet d'obtenir une description suffisamment exhaustive des mots à reconnaitre utilisable dans une approche markovienne.

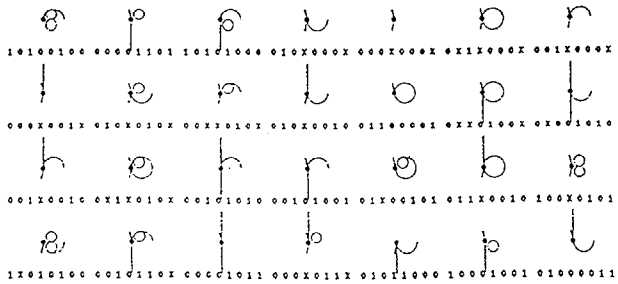

figure 4 : quelques exemples de l'alphabet sélectionné.

3. MODELISATION STOCHASTIQUE

Les deux modèles structurels présentés au paragraphe 2 sont utilisés pour construire deux modèles de Markov non cachés dont les états sont soit les traits, soit les graphèmes de l'alphabet sélectionné. L'ordre de ces modèles est évalué en utilisant un critère d'information d'Akaïke. Un problème fréquemment rencontré lorsqu'on utilise un modèle Markovien est la décroissance de la probabilité évaluée en fonction de la longueur de la séquence analysée. Nous proposons dans la

deuxième partie de ce paragraphe un moyen de prendre en compte la longueur de la séquence lors de l'évaluation de la probabilité du modèle. La méthode proposée est basée sur une modèlisation de chaque mot du dictionnaire, ceci n'est bien sûr envisageable que pour un vocabulaire de petite taille tel que celui utilisé pour exprimer un montant littéral.

3.1 Choix de l'ordre du modèle

L'information fournie par chacun des deux modèles structurels, traits ou graphèmes, est directement utilisée pour définir les états de chacun des modèles de Markov. Nous utilisons donc des modèles non cachés (MM) définis par: - le nombre des états possibles N (N= 12 pour le modèle traits, N = 42 pour le modèle graphèmes). - un vecteur de probabilités initiales Π (de dimension 12 ou 42). - une matrice de transition A d'ordre c où c désigne l'ordre du modèle. En 1973, H. Akaïke a proposé un critère permettant d'évaluer l'ordre optimal d'un modèle auto-régressif en utilisant un coût du type contrast de Kullback. De ce fait, le modèle le plus efficace permettant de prédir une observation n'est pas nécessairement celui le plus riche en paramètres. C'est même généralement le contraire qui est observé, puisqu'au-delà d'un ordre c optimal, l'ajout de paramètres sur le modèle conduit à une perte d'efficacité du modèle. Un critère général peut s'écrire comme suit:

$$A(c) = 2\alpha(c) - \frac{2}{n}\sum_{i=1}^{n} \log\left[f(\theta_{n,c}, X_i)\right],$$

où $f(\theta_{n,c},X_i)$ est la densité de probabilité de $\theta_{n,c}$ l'estimateur du processus aléatoire inconnu θ dépendant de c paramètres; $(X_1,X_2,...,X_n)$ sont les observations de ce processus aléatoire. Le terme $\alpha(c)$ est un terme de pondération dont l'expression varie selon les auteurs. Dans l'étude présentée nous avons choisi d'utiliser le critère d'Akaïke standard car l'ordre optimal recherché est généralement petit et ne conduit donc pas à une sur-paramétrisation importante du modèle, ce qui est le cas de façon général pour ce critère. De plus, comme le montre l'étude présentée dans [6], la complexité des calculs conduit à utiliser des critère aussi simples que possible.
Notons $\theta_{n,m}$ le processus de Markov d'ordre m (m>c) modèlisant la suite d'observations $Q=(X_1,X_2,...,X_n)$. Si l'on désigne par $\eta_{c,m}$ l'approximation de l'espérance du contraste de Kullback entre $\theta_{n,c}$ et $\theta_{n,m}$, $\eta_{c,m}$ et s'écrit alors:

$$\eta_{c,m} = -\frac{2}{n}\sum_{i=1}^{n} \log\left[\frac{f(\theta_{n,c}, X_i)}{f(\theta_{n,m}, X_i)}\right],$$

$\eta_{c,m}$ peut être calculé récursivement et suit une loi du Chi-2. On obtient alors le critère MAICE défini par:

$$MAICE(c) = -2DF(\eta_{c,m}) + \eta_{c,m}$$

où DF désigne le degré de liberté. Dans notre cas, ce terme représente la pénalisation $\alpha(c)$. L'ordre optimal c du Modèle de Markov est donc celui qui minimise l'expression du critère MAICE. On choisi donc un ordre initial m choisi *a priori* supérieur à c. En pratique, des valeurs importantes de m nécessitent des calculs de complexité importante, tandis que des valeurs de m choisies trop faibles ne permettent pas de déterminer correctement c.

Applications:

L'ordre optimal déterminé par MAICE sur la base de données dont nous disposons de 6000 mots est c=2, en débutant la recherche à partir de m=5, ceci pour le modèle structurel graphème. Ce résultat justifie donc, au sens du contraste de Kullback, d'utiliser un modèle d'ordre deux et donc d'effectuer un apprentissage des tri-grammes de graphèmes rencontrés sur la base au lieu d'effectuer *a priori* l'apprentissage sur l'ensemble des séquences rencontrées sur cette même base. L'évaluation du critère MAICE pour m allant de 1 à 5, effectuée pour le modèle structurel en traits permet de déterminer l'odre optimal pour c=2. Dans la suite de cette discussion les Modèle de Markov seront choisi d'ordre deux, qu'il s'agisse du modèle trait ou graphème.

3.2 Classifieurs Markoviens

Le classifieur mot, de type bayesien, doit être capable d'identifier une séquence d'observations $Q=(X_1,X_2,...,X_n)$ en déterminant la probabilité de chacun des modèles considérés conditionnellement à cette séquence d'observations indépendemment de la longueur de cette séquence. Cette probabilité s'exprime alors pour un modèle d'ordre deux par:

$$P(M_i / Q) = \frac{P(Q / M_i)P(M_i)}{P(Q)} = \frac{P(M_i)P(X_1 / M_i)P(X_2 / X_1, M_i)P(X_3 / X_1, X_2, M_i)...}{P(X_1)P(X_2 / X_1)P(X_3 / X_1, X_2)...} \; x \frac{...P(X_n / X_{n-2}, X_{n-1}, M_i)}{...P(X_n / X_{n-2}, X_{n-1})}$$

en supposant que la séquence peut être modèlisée par un modèle d'ordre deux (c=2). Cette expression peut être résumée par l'expression suivante:

$$P(M_i / Q) = a_0 P(M_i) \prod_{t=3}^{n} \frac{T(t, M_i)}{\sum_j T(t, M_i) P(M_j / Q_{t-1})}$$

où
- α_0 est un facteur dépendant des conditions initiales pour t=1 et t=2. - $Q_t=(X_1,...,X_t)$ et $Q_n = Q$.
- $T(t,M_j)$ désigne la matrice de transition entre états à l'ordre deux, soit $P(X_t/X_{t-1},X_{t-2},M_j)$ pour le modèle M_j.
De cette façon on peut déterminer de façon dynamique $P(M_i/Q)$ en évaluant $P(M_i/Q_t)$ par itérations successives:
pour t = 3 à t = n:

$$P(M_i/Q_t) = P(M_i / Q_{t-1}) \frac{T(t, M_i)}{\sum_j T(t, M_i) P(M_j / Q_{t-1})}.$$

Le dénominateur de cette expression est le même pour tous les modèles et permet ainsi d'évaluer une expression indépendante de la longueur de la chaîne. La probabilité initiale $P(M_i)$ peut être considérée de densité uniforme, ou bien déterminée sur la base d'apprentissage en évaluant la fréquence statistique de chaque mot. Dans un premier temps nous avons fait l'hypothèse d'une équi-probabilité des mots sur la base mais l'examen des fréquences statistiques de chaque mot à montré qu'il est préférable d'utiliser la seconde approche. Ceci se justifie simplement par le fait que le mot "francs" est largement plus représenté que tous les autres mots; au contraire, le mot "zéro" est très rare. Finalement nous proposons une liste de mots

ordonnée pour chacune des chaînes de traits ou de graphèmes observées.

4. RESULTATS

Chaque mot a été modélisé par un modèle de Markov traits ou graphèmes. Notre base de données est composée d'environ 7200 mots appartenant à 27 classes différentes et divisée en deux sous-bases, 60% pour la base d'apprentissage, 40% pour la base de test. Les tableaux ci-dessous donnent les taux de bonne reconnaissance en fonction du rang dans la liste de proposition.

Modèle traits

Rang dans la liste	R(1)	R(2)	R(3)	R(4)	R(5)
Reco (%)	21	35	46,5	56	63,2

Modèle de Markov d'ordre 1 sur les traits.

Rang dans la liste	R(1)	R(2)	R(3)	R(4)	R(5)
recognition (%)	34	52,9	64,8	73,2	78,8

Modèle de Markov d'ordre 2 sur les traits.

Modèle graphèmes

Rang dans la liste	R(1)	R(2)	R(3)	R(4)	R(5)
recognition (%)	43,9	59,8	69,4	75	80,2

Modèle de Markov d'ordre 1 sur les graphèmes.

Rang dans la liste	R(1)	R(2)	R(3)	R(4)	R(5)
recognition (%)	69,6	80	86,2	90,1	92,4

Modèle de Markov d'ordre 2 sur les graphèmes.

Ces tableaux confirment les résultats théoriques obtenus par le critère MAICE. Ceci indique que pour l'alphabet de 42 graphèmes sélectionné, c'est le choix de l'alphabet de graphèmes, et la méthode de sélection utilisée qui devrait être remise en cause. Dans tous les cas, ces résultats semblent intéressants car ils s'appuyent sur une description structurelle relativement pauvre (8 traits élémentaires différents).
Une seconde voie consisterait à fusionner les résultats des deux classifieurs.

5.CONCLUSION

L'étude présentée dans cette communication aborde le problème de l'optimisation de Modèles Markoviens par l'utilisation de critère d'information dans le cadre de la reconnaissance de mots manuscrits cursifs. La méthode proposée a été utilisée pour optimiser deux modèles structurels construits autour des traits ou des graphèmes. Les résultats montrent la bonne adéquation entre les résultats théoriques concernant l'odre du modèle et les taux de reconnaissance pour ce même ordre. Ces résultats montrent également la pertinence de la description structurelle utilisée pour la reconnaissance de mots très mal écrits. L'ajout de caractéristiques supplémentaires à l'ensemble des 8 traits utilisés peut être envisagé. Comme nous l'avons indiqué, les représentations structurelles utilisées ne permettent pas, de la façon dont elles sont utilisées pour cette application, une extension à un vocabulaire de plus grande taille, nous envisageons actuellement d'utiliser cette information structurelle pour construire des modèles basés sur une description explicite des mots en lettres de façon à travailler sur des lexiques de plus grande taille.

BIBLIOGRAPHIE

[1] BOZINOVIC R-M, SRIHARI S-N: *"Off-line cursive word Recognition"* - IEEE- Trans. PAMI, **11**, 1, pp 68-83, Jan 1989.

[2] G. LORETTE, Y. LECOURTIER, "Reconnaissance et interprétation de textes manuscrits hors-ligne", BIGRE, 80, pp. 109-135, 1992.

[3] RABINER R.: *"A Tutorial on HMM and selected Applications in Speech Recognition"* - Proc. IEEE, **77**, 2, pp 257-284, 1989.

[4] GILLOUX M.: *"Research into the new generation of character and mailing address recognition systems at the french post office research center"*,- Pattern Recognition Letters, **14**, pp 267-276, 1993.

[5] CHEN M-Y, KUNDU A., ZHOU J.: *"Off-line handwritten word Recognition using a hidden Markov Model type stochastic Network"* - IEEE Trans. PAMI, **16**, 5, pp 481-496, May 1994.

[6] KUNDU A., HE Y.: *"On optimal Order in modeling Sequence of Letters in Words of common Language as a Markov Chain"* - Pattern Recognition, **24**, 7, pp 603-608, 1991.

[7] PAQUET T., LECOURTIER Y.: *"Recognition of handwritten Sentences using a restricted Lexicon"* - Pattern Recognition, **26**, 3, pp 391-407, 1993.

[8] AKAIKE H.: *"A new Look at statistical model Identification"* - IEEE Trans. ACVA, **19**, 6, pp 716-723, 1974.

[9] RISSANEN J.: *" Stochastic Complexity in Statistical Inquiry"* - World Scientific Series in Computer Science, 15, 178 p, 1989.

[10] M. AVILA, C. OLIVIER, PAQUET T., Y. LECOURTIER: *"Procédure de reconnaisssance de l'écriture manuscrite basée sur des chaînes de Markov cachées et appliquée à un vocabulaire limité"*, 14ème colloque GRETSI, 2, 803-806, juan les Pins, 13-16 septembre 1993.

[11] A. BELAID, Y. BELAID, *"Reconnaissance des formes, méthodes et applications"*, InterEditions (Paris), 1992.

[12] C. OLIVIER, T. PAQUET, M. AVILA, Y. LECOURTIER, *"Recognition of Handwritten Words Using Stochastic Models"*, ICDAR'95 Montréal Aug. 1995.